WILEY-BLACKWELL

Cladistics 29 (2013) 643–662

Cladistics

10.1111/cla.12013

# Uncovering historical signature of mitochondrial DNA hidden in the nuclear genome: the biogeography of *Schistocerca* revisited

Hojun Song[a,b,*], Matthew J. Moulton[b], Kevin D. Hiatt[b] and Michael F. Whiting[b]

[a]*Department of Biology, University of Central Florida, Orlando, FL, 32816, USA;* [b]*Department of Biology and M. L. Bean Life Science Museum, Brigham Young University, Provo, UT, 84602, USA*

Accepted 17 December 2012

**Abstract**

Inadvertent coamplification of nuclear mitochondrial pseudogenes (numts) is a serious problem in mitochondrial systematics, but numts can also be a valuable source of information because they represent ancient forms of mtDNA. We present a conceptual framework of numt accumulation, which states that in a given species there can be two types of numts, synaponumts and autaponumts, resulting from integration occurring respectively before and after a speciation event. In a given clade, a species that diverged early can only have its own autaponumts as well as synaponumts that were already present in the genome of the last common ancestor. A species that diverged more recently may, however, have many different synaponumts integrated at each different divergence as well as its own autaponumts. Therefore it is possible to decipher the evolutionary history of a species based on the phylogenetic distribution of numts in a simultaneous analysis of numts and extant mtDNA. In this study, we test this idea empirically in the context of addressing a controversial question regarding the biogeography of the grasshopper genus *Schistocerca* Stål (Orthoptera: Acrididae), based on numts of the cytochrome *c* oxidase subunit I (COI) gene. We find that our empirical data can be explained adequately by our conceptual framework, and that the phylogenetic distribution of COI numts reveals intricate evolutionary histories about past speciation events that are otherwise difficult to detect using conventional markers. Our study strongly favours the Old World origin of the desert locust, *Schistocerca gregaria* and the New World *Schistocerca* species are descendants from an ancestral *gregaria*-like species that colonized the New World via westward transatlantic flight. However, the phylogenetic distribution of *S. gregaria* numts raises a distinct possibility that there might have been multiple founding events from Africa to America to give rise to the present-day diversity of the genus. This is a case study for a creative use of numts as molecular fossils, and we demonstrate that numts provide an interesting and powerful phylogenetic signal, much more than what extant mtDNA or nuclear gene sequences might be able to provide.



Nuclear mitochondrial pseudogenes (numts) are fragments of the mitochondrial genome that have been integrated into the nuclear genome (Lopez et al., 1994; Bensasson et al., 2001a; Hazkani-Covo et al., 2010). Numts are extremely prevalent across eukaryotes (Bensasson et al., 2001a; Richly and Leister, 2004; Hazkani-Covo et al., 2010) and can be inadvertently coamplified using a PCR-based method (Collura and Stewart, 1995; Zhang and Hewitt, 1996; Sorenson and Quinn, 1998; Bensasson et al., 2001a; Thalmann et al., 2004; Benesh et al., 2006; Sword et al., 2007; Song et al., 2008; Buhay, 2009; Moulton et al., 2010). Without a proper screening, they can potentially compromise the orthology assumption of nucleotide characters in many areas of mitochondrial systematics, including phylogeography (Sorenson and Quinn, 1998; Thalmann et al., 2005; Podnar et al., 2007; Hlaing et al., 2009; Baldo et al., 2011; Bertheau et al., 2011); phylogenetics (Arctander, 1995; van der Kuyl et al., 1995; Sunnucks and Hales, 1996; Lopez et al., 1997; Mirol et al., 2000; Williams and Knowlton, 2001; Benesh et al., 2006; Hazkani-Covo, 2009; Berthier et al., 2011); molecular ecology (Triant and Hayes, 2011); and DNA barcoding (Song et al., 2008; Buhay, 2009; Moulton et al., 2010).

*Corresponding author:
*E-mail address*: song@ucf.edu

While many studies have focused on the adverse effect of numts, numts have interesting properties that potentially can be used to infer useful information about past evolutionary events (Zischler et al., 1995; Perna and Kocher, 1996; Zischler, 2000; Bensasson et al., 2001a; Hazkani-Covo, 2009; Jensen-Seaman et al., 2009; Baldo et al., 2011; Miraldo et al., 2012). Two lines of evidence suggest that numts carry an important historical signature. First, the fact that many divergent numts can be recovered from a single individual implies that the nuclear genome harbours the fragments of ancient mtDNA incorporated at different times throughout the evolutionary history of the species (Zischler et al., 1995; Perna and Kocher, 1996; Bensasson et al., 2000, 2001a; Song et al., 2008). Second, the fact that numts can easily be coamplified with conserved primers (Perna and Kocher, 1996; Thalmann et al., 2004; Benesh et al., 2006; Sword et al., 2007; Song et al., 2008; Moulton et al., 2010), and that the mutation rate in the nuclear genome is generally much slower than in the mitochondrial genome (Brown et al., 1982; Fukuda et al., 1985; Arctander, 1995; Lopez et al., 1997; Schmitz et al., 2005), supports the idea that numts have remained relatively intact and accumulated few mutations since the time of nuclear integration. In other words, it is possible to consider numts to be molecular fossils of the ancestral mtDNA that have been lodged silently in the nuclear genome (Zischler et al., 1995; Perna and Kocher, 1996; Zhang and Hewitt, 1996; Lopez et al., 1997; Bensasson et al., 2001a; Thalmann et al., 2005; Hazkani-Covo, 2009).

A number of researchers have developed various methods of using numts as phylogenetically informative markers. Numts have been used as molecular outgroups for geese (Ruokonen et al., 2000), humans (Zischler et al., 1995), and tuarata (Hay et al., 2004). This approach is especially useful for taxa with no appropriate extant outgroup. Mutational events associated with the nuclear integration of mtDNA and their presence/absence status in different species could also be informative when inferring phylogeny (Zischler, 2000). Hazkani-Covo (2009) demonstrated this idea empirically by analysing shared numt insertions in five fully sequenced primate genomes. For example, she showed that humans and chimpanzees share some numt insertions integrated in their common ancestor and that these numts are not present in other primates. The time of nuclear integration of mtDNA can also be calculated to estimate the divergence time between two taxa (Lopez et al., 1997; Thalmann et al., 2004). Below the level of species, numts have been used to infer past mitochondrial diversity and gene exchange among different mitochondrial lineages (Baldo et al., 2011; Miraldo et al., 2012).

Building on these previous studies, we have developed the following conceptual framework of numt accumulation, which can help decipher evolutionary relationships of closely related species (Fig. 1). Figure 1a depicts a hypothetical phylogeny of four closely related species based on mtDNA. Throughout the diversification of these four lineages, nuclear integration of mtDNA occurs continuously, which results in numt accumulation in the nuclear genome before and after speciation events of each species. There would be nuclear integration events that happen after the point of species divergence to the present ($N_{A1}$, $N_{A2}$, $N_{A3}$…), and the resulting numts would be similar to the extant mtDNA of that particular species (species

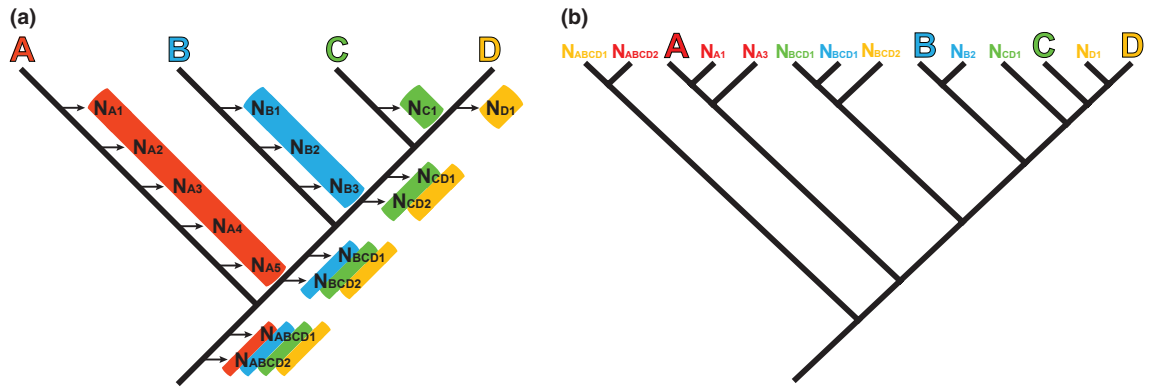


Fig. 1. A conceptual framework of numt accumulation. (a) Throughout the diversification of species A–D, fragments of mtDNA from these four species can be integrated continuously in the nuclear genome, resulting in numts. There can be two types of numts: autaponumts, which result from nuclear integration events that happen after the point of species divergence to the present ($N_{A1}$, $N_{B1}$, $N_{C1}$, $N_{D1}$…) and synaponumts, which result from nuclear integration in the most recent common ancestor (MRCA) prior to each species divergence ($N_{ABCD1}$, $N_{BCD1}$, $N_{CD1}$…). For each type there can be multiple nuclear integration events. Background colour of each numt indicates the species from which the numt can be coamplified. (b) Some numts can be coamplified with the mtDNA using conventional PCR techniques. If the numts and mtDNA are analysed simultaneously in a phylogenetic study, autaponumts form a monophyletic clade with the respective ortholog, while synaponumts from species that do not share MRCA ($N_{ABCD1}$ from species A and $N_{ABCD1}$ from species D) could form a monophyletic clade. See text for detail.

A) and present in the nuclear genome of only that species. Because these numts are unique to one lineage, we refer to them as "autaponumts". An autaponumt is similar to the cladistic concept of an autapomorphic character in the sense that it is derived after speciation and contains no phylogenetic information for resolving relationships in a phylogenetic analysis. There would also be nuclear integration events that occur prior to speciation events throughout the diversification of a group, which we refer to as "synaponumts". A synaponumt is similar to a synapomorphy in the sense that the shared possession of the same numt among multiple extant taxa provides evidence for grouping those taxa into a single clade. Hazkani-Covo (2009) demonstrated empirically the presence of synaponumts in primates, although she did not use this specific terminology. Effectively, a species that diverged earlier (species A) would possess only two types of numt: synaponumts resulting from the nuclear integration in the oldest common ancestor ($N_{ABCD1}$), which may be called symplesionumts, similar to the concept of a symplesiomorphy, and its own autaponumts. On the other hand, a recently diverged species (species D) could possess many different synaponumts that have accumulated between each divergence event as well as its own autaponumts. Both autaponumts and synaponumts are subject to duplications (Lopez et al., 1994; Bensasson et al., 2003; Hazkani-Covo et al., 2003; Behura, 2007; Triant and DeWoody, 2007), which may result in high density of similar numts, but each unique nuclear integration event would yield numts with distinct sequence characteristics at the time of integration (Bensasson et al., 2001a).

Using a PCR-based approach, it is possible to easily coamplify and sequence numts (Song et al., 2008). If the randomly coamplified numts and the extant mtDNA were analysed together in a phylogenetic framework, we would expect to recover a topology such as presented in Fig. 1b. The autaponumts would always form a monophyletic clade with the extant mtDNA of that species (A + $N_{A1}$, B + $N_{B2}$). The synaponumts accumulated between each divergence event may be amplified from different species and those numts can form a monophyletic group on their own [$N_{BCD1}$ amplified from species B (blue) and $N_{BCD1}$ amplified from species C (green)]. The numts resulting from duplications would form a monophyletic group with the original numt resulting from each unique nuclear integration event. The numts of the early branching species can only be recovered at the base of tree, while the numts of the recently diverged species can be recovered all along the backbone of the tree. For example in Fig. 1b, the numts of the early branching species A would group either with the mtDNA of the same species (A + $N_{A1}$) or with the ancient synaponumt from other species [$N_{ABCD1}$ from species D (yellow) and $N_{ABCD2}$ from species A (red)]. However, they would never be placed apically in the tree past the divergence of species A. The numts of the recently diverged species D would form monophyletic groups with the synaponumts of many other species because there would have been a higher chance for the ancient mtDNA of common ancestors at different times to be integrated into the nuclear genome. If this conceptual framework of numt accumulation is found to explain empirical data adequately, we can deduce phylogenetic placements of extant taxa from the phylogenetic distribution of their numts based on a simultaneous analysis of the extant mtDNA and their numts. This approach would be especially useful for recently diverged groups, which often do not have reliable molecular loci to determine a correct phylogeny (Maddison and Knowles, 2006; Knowles and Carstens, 2007).

In this study, we conduct a thorough survey of numts of the cytochrome *c* oxidase subunit I (COI) gene in the grasshopper genus *Schistocerca* Stål (Orthoptera: Acrididae: Cyrtacanthacridinae), which is known to include a species with an unusually high number of numts (Song et al., 2008). Using the numts generated from the multiple species in a single genus, we demonstrate empirically the utility of the proposed conceptual framework of numt accumulation (Fig. 1). The use of numts as phylogenetic markers is especially suitable for *Schistocerca*, which is a recently diverged genus (Lovejoy et al., 2006) and displays an unusual biogeographical distribution (Amédégnato, 1993; Song, 2004a; Lovejoy et al., 2006). The genus contains about 50 species, but only one species, the desert locust *S. gregaria*, occurs in the Old World, while the rest of the genus occur in the New World (Dirsh, 1974). Regarding the biogeography, two main alternative hypotheses have been proposed, which centre on the phylogenetic placement of *S. gregaria* within *Schistocerca*. The first hypothesis, referred to as the "New World origin of the desert locust hypothesis", states that the ancestral desert locust originated from the New World and flew across the Atlantic Ocean to give rise to the present-day desert locust in Africa and the Middle East. This hypothesis can be supported if the desert locust is phylogenetically nested deep within the New World species. Song (2004a) supported this hypothesis based on a morphological phylogeny that placed *S. gregaria* near the Americana Complex *sensu* Harvey (1981), which also includes other swarming locust species in the genus. The second hypothesis, the "Old World origin of the desert locust hypothesis", states that the New World *Schistocerca* species are descendants of a "*gregaria*-like" ancestor from the Old World that crossed the Atlantic Ocean once. This hypothesis can be supported if the desert locust is placed basally to the rest of the New World *Schistocerca*. Based on mitochondrial ribosomal RNA genes,

Lovejoy et al. (2006) reconstructed a phylogeny of *Schistocerca* and placed the desert locust at the base of the phylogeny with strong support values, suggesting that the diversity of *Schistocerca* in the New World must have been a result of a single colonization event via westward transatlantic flight from Africa to America by an ancestral desert locust followed by an explosive radiation. To date, there is yet to be an independent test of the biogeographical hypotheses of *Schistocerca* to resolve the conflict between Song (2004a) and Lovejoy et al. (2006), therefore a numt-based analysis might be able to shed a new light on the biogeography of *Schistocerca*.

## Material and methods

### *Taxon sampling*

In order to study numt accumulation among closely related species, we included 20 species of *Schistocerca*, one species of *Halmenus* Scudder, and five outgroup taxa belonging to Cyrtacanthacridinae that have been shown in previous studies to be appropriate outgroups (Song and Wenzel, 2008) (Table 1). Specifically, we included *Halmenus* because its phylogenetic position relative to *Schistocerca* is unclear. Dirsh (1974) and Song (2004a) considered it a sister genus to *Schistocerca* based on morphological evidence, but a recent molecular study suggested that it might belong to *Schistocerca* (Lovejoy et al., 2006). Although our taxon sampling covered less than 50% of the known diversity of *Schistocerca*, the species included represented evenly all major species groups within the genus from all their geographical ranges (Song, 2004a). Therefore our taxon sampling was appropriate for the questions addressed in this study.

Because typical numts are known to be relatively short, less than 1000 bp in size (Richly and Leister, 2004), we performed long-range PCR reactions to target only the mtDNA and avoid numts using the primers, LM.ND2F (5′-GCC ACC AAT ACT AGG ATT TCT ACC-3′) and LM.ND3R (5′-AGG GTC AAA TCC ACA TTC AAA TGG-3′). These PCR products, ca. 4.5 kb in size and spanning from the ND2 to the ND3 gene, were used as templates for nested PCR reactions using Folmer primers (Folmer et al., 1994) to obtain the mtDNA-encoded COI genes. For *S. gregaria* in particular, we extracted the Folmer region from the published mitochondrial genome (GenBank accession: GQ491031, Erler et al., 2010) to ensure orthology. To obtain the numts, we followed the protocol described in Song et al. (2008). In short, we performed PCR reactions using the Folmer primers from genomic extracts, cloned the PCR products, picked about 60 colonies per taxon, and sequenced the clones. We generated a total of 1001 cloned sequences, many of which were identical to the corresponding mtDNA sequences. There were also a number of clones represented by multiple copies. After accounting for the duplicates, we obtained a total of 629 unique cloned sequences across 21 ingroup taxa. In all PCR reactions for numt generation, we used Elongase Enzyme mix (Invitrogen Corporation, Carlsbad, CA, USA) because of its high fidelity and low error rate (0.015% or 0.0987 bp per 658-bp Folmer region) (Leroux et al., 1997).

In addition to the COI gene, we generated internal transcribed spacer between 5.8S rRNA and 28S rRNA (ITS2), histone 3 (H3), and elongation factor 1α (EF-1α) for all taxa to reconstruct a robust molecular phylogeny of *Schistocerca*. We followed the PCR protocols described in Collins and Paskewitz (1996) and Svenson and Whiting (2004) for generating ITS2 and H3, respectively. For EF-1α, we designed SEF1F (5′-CTG GAT GGC ATG GCG ACA ACA TGC-3′) and SEF1R (5′-AAG TGC AGT TCA CTC ACC AGT ACG-3′) to generate the sequences. For all nuclear genes, we used Platinum Taq polymerase (Invitrogen Corporation, Carlsbad, CA, USA) following the manufacturer's recommendation.

The resulting sequences were deposited at GenBank with accession numbers GU115845–GU116477 and JQ796285–JQ796365. A complete list of taxa included in this study is shown in Table 1.

### *Sequence characterization*

We performed a BLAST search (http://blast.ncbi.nlm.nih.gov) on each cloned sequence against the known nucleotide sequence collection (nr/nt) using the MEGABLAST option to ensure that we only included the sequences of mtDNA in origin in the analysis. In order to determine the number of insertions and deletions, each cloned sequence was compared against the reference mtDNA of the corresponding species by performing a series of pairwise alignments using MUSCLE (Edgar, 2004) under default parameters. We also inferred the number of stop codons by translating each cloned sequence to amino acid sequence using the invertebrate genetic code. The inferred indels and stop codons were treated as intrinsic features of each numt sequence.

Bensasson et al. (2001a) suggested that numts might have a different nucleotide composition compared with their mitochondrial counterpart because the nuclear genome would be evolving under different conditions from the mitochondrial genome. If this were the case, we would expect to find a different pattern of nucleotide evolution in numts relative to the mtDNA. To investigate this pattern in our dataset, we trimmed the master alignment (see below) to construct an individ-

Table 1
A list of taxa and collecting information used in this study

| Species | Collecting locality | Collecting date | Collector |
|---|---|---|---|
| *Anacridium* sp. | Namibia: Etosha NP Devilwater Spring (S18″59.266′ E15″15.593′) | 18-V-2004 | K.B. Miller |
| *Acanthacris ruficornis* (Fabricius) | Namibia: Etosha NP (S19″10.484′ E15″54.872′) Ojivisandy | 15-22-V-2004 | K.B. Miller |
| *Cyrtacanthacris tatarica* (Linneaus) | South Africa: Kruger NP between Ngotso Dam and Satara | 16-III-2005 | G. Svenson, S. Cameron& S. Bybee |
| *Valanga* sp. | PNG: Eastern Highlands Province Herowana Village Crater Mountain Research Area | 15-19-VII-2001 | S. Bradler, K. Jarvis & G. Svenson |
| *Austracris guttulosa* (Walker) | PNG: ENBP Kulau Lodge (S04″11.990′ E152″5.984′) | 21-29-VII-2008 | M.F. Whiting |
| *Halmenus robustus* Scudder | Ecuador: Galapagos Islands: Isla Santa Cruz, trail to Tortuga Bay | 7-III-1996 | D. Otte |
| *Schistocerca albolineata* (Thomas) | USA: AZ, Pima Co. Tucson Mts. off Gates Pass Rd.(N32″13.366′ W111″05.570′) | 3-IX-2002 | H. Song |
| *Schistocerca shoshone* (Thomas) | USA: AZ, Yuma Co. Tacna, Jojoba plantation on 36 Ave. ext 37 | 4-XI-2002 | H. Song |
| *Schistocerca ceratiola* Hubbell & Walker | USA: FL, Putnam Co. Katherine Ordway Preserve (N29″41.541′ W081″58.657′) | 7-VIII-2002 | H. Song |
| *Schistocerca alutacea* (Harris) | USA: FL, Walton Co. DeFuniak Springs, off SR 90 W Oakwood Hills (N30″45.418′ W086″16.661′) | 8-VIII-2002 | H. Song |
| *Schistocerca rubiginosa* (Harris) | USA: FL, Walton Co. DeFuniak Springs, off SR 90 W Oakwood Hills (N30″45.418′ W086″16.661′) | 8-VIII-2002 | H. Song |
| *Schistocerca damnifica* (Saussure) | USA: FL, Levy Co. Off SR 121 (N 29″25.908′ W082″24.060′) | 6-VIII-2002 | H. Song |
| *Schistocerca obscura* (Fabricius) | USA: FL, Marion Co. Jct. CR 318 & CR 329 (N 29″24.480′ W082″17.775′) | 4-VIII-2002 | H. Song |
| *Schistocerca americana* (Drury) | USA: FL, Marion Co. Jct. CR 318 & CR 329 (N 29″24.480′ W082″17.775′) | 4-VIII-2002 | H. Song |
| *Schistocera vaga* (Scudder)* | USA: AZ, Pima Co. Tucson | 4-XI-2002 | R.F. Chapman |
| *Schistocerca lineata* Scudder | USA: OK, Comanche Co. Fort Sill Military Reservation West Range (N34″39.517′ W98″33.054′) | 18-VII-2004 | H. Song |
| *Schistocerca pallens* (Thunberg) | Mexico: Rancho El Gualul Motorway Estacion, Manuel a Ebano S.LP. Km 11 | 10-XI-2001 | L. Barrientos-Lozano |
| *Schistocerca piceifrons* (Walker) | Mexico: Rancho El Gualul Motorway Estacion, Manuel a Ebano S.LP. Km 11 | 27-XI-2001 | L. Barrientos-Lozano |
| *Schistocerca impleta* (Walker)* | Mexico: Rancho El Gualul Motorway Estacion, Manuel a Ebano S.LP. Km 11 | 10-XI-2001 | L. Barrientos-Lozano |
| *Schistocerca literosa* (Walker) | Ecuador: Galapagos Islands, Espanola | 2-V-1992 | S.B. Peck |
| *Schistocerca melanocera* (Stål) | Ecuador: Galapagos Islands, Isla Floreana, Finca Cruz ft. of Cerro Communista | 28-III-1996 | D. Otte |
| *Schistocerca gregaria* (Forskål) | Oxford lab culture (provided by S. Simpson) | | S.J. Simpson |
| *Schistocerca flavofasciata* (De Geer) | British Virgin Is: Guana I | 9~15-X-2002 | B. Valentine |
| *Schistocerca quisqueya* Rehn & Hebard | Dominican Republic: RD-060, Monte Rio, ~4 km W of beach, ~70 m (N18″22.972′ W70″43.036′) | 17-XI-2002 | D. Perez-Gelabert |
| *Schistocerca socorro* (Dirsh) | Mexico: Islas Revillagigedo, Isla Socorro (N18″51.527′ W110″59.249′) | 20~27-X-2004 | H. Song |
| *Schistocerca cancellata* (Serville) | Nonato (S08″50.407′ W042″33.778′) | VIII-2005 | S. Bybee & K. Dittmar |

*Dirsh (1974) synonymized 16 names under *Schistocerca nitens* (Thunberg) without any justification, but taxonomists generally agree that most of the synonymized species are actually valid species. In this study, we included two of such species and used their names prior to the synonymy.

ual alignment for each species containing the reference mtDNA and all cloned sequences for the particular species. We then used the matched pairs test of symmetry (Ababneh et al., 2006), as implemented in SeqVis (Ho et al., 2006), which tests against the null hypothesis that a pair of sequences (e.g. mtDNA and a cloned sequence) has evolved under the same stationary, reversible, and homogeneous conditions. Additionally, we also calculated AT% for all sequences to compare the base composition between the reference mtDNA and the cloned sequences.

Finally, we calculated pairwise distances (uncorrected *p*-distance) in MEGA 5.05 (Tamura et al., 2011) between cloned sequences and the reference mtDNA to estimate nucleotide sequence divergence.

### *Phylogenetic analyses*

The cloned sequences varied considerably in terms of the length, the presence of indels and stop codons, base composition, and nucleotide sequence divergence compared with the corresponding reference mtDNA. To explore how the cloned sequences were phylogenetically related to the respective mtDNA sequences, we first constructed a single master dataset consisting of all 629 unique clones, 21 ingroup orthologs and five outgroup sequences (a total of 655 terminals). This dataset was aligned in MUSCLE using default parameters, which resulted in 769 aligned nucleotide characters. According to the conceptual framework of numt accumulation presented in the introduction, the phylogenetic placement of numts of any given species could provide a clue on the divergence of that species. To explore this idea further, we created 21 separate species-specific datasets from the master alignment that include all cloned sequences of a given species and the reference mtDNA sequences for all ingroup and outgroup taxa.

These datasets were analysed in a parsimony framework in TNT (Goloboff et al., 2003) using New Technology Search options with gaps treated as missing. For each dataset, we first reset RAM to 1000 Mbytes, held 100 000 trees, and set *Anacridium* as a root prior to tree search. The minimum length was initially searched 10 times using ratchet (Nixon, 1999), sectorial search, drift and tree fusing (Goloboff, 1999) under default parameters and by setting random seed as 0 and the initial driven search level at 15. Then the driven search level was increased to 99 to search more thoroughly for the minimum length. Once the minimum length was identified, we set to give up search at the particular minimum length score, and searched for the most parsimonious trees for 1000 times. Both Bremer support and bootstrap support values were calculated to assess nodal support. The resulting tree files from TNT were manually converted to Newick format to view them in Dendroscope (ver. 2.7.4) (Huson et al., 2007), which was also used to calculate strict consensus. The presence of insertions, deletions, and stop codons was mapped onto the resulting strict consensus trees, which were illustrated in iTOL (Letunic and Bork, 2007, 2011).

Additionally, we created two alignments of 26 terminals (21 ingroups and five outgroups). The first was a COI dataset (658 aligned nucleotide characters) to estimate an mtDNA gene tree. The second was a four-gene dataset (2128 aligned nucleotide characters) consisting of COI, ITS2, H3, and EF-1α genes to estimate a species tree. For the four-gene dataset, individual genes were aligned separately in MUSCLE using default parameters and concatenated into a single matrix using MacClade 4.08 (Maddison and Maddison, 2005). These two datasets were analysed in both parsimony and Bayesian frameworks. For the four-gene dataset, we performed a mixed-model partitioned Bayesian analysis using MrBayes 3.1.1 (Ronquist and Huelsenbeck, 2003). We first selected the best-fit model for each gene partition as recommended by MrModelTest (Nylander, 2004) and ran four runs with four chains each for 20 million generations, sampling every 1000 generations. We plotted the likelihood trace for each run to assess convergence, and discarded an average of 25% of each run as burn-in.

## Results

### *Sequence characterization*

We sequenced a total of 629 unique clones from the PCR products generated using the Folmer primers from all 21 ingroup species. Of the 629 clones, we found 214 sequences that had accumulated stop codons, 70 with insertions, and 206 with deletions, which helped identify them clearly as numts (Table 2). Deletions appeared to occur more frequently than insertions. Several clones had only stop codons or indels, suggesting that these mutations occurred randomly in the nuclear genome. The clones with many stop codons also had indels, suggesting that indel events might have caused frameshift mutations, resulting in stop codons. Many clones with numerous stop codons had very low sequence divergence from the reference mtDNA (Fig. 2b), suggesting that the stop codons could be easily accumulated in recent numts. The amount of nucleotide divergence varied greatly across taxa (Table 3), but the majority of the clones had less than 3% sequence divergence from the reference mtDNA (405 clones, or 64.4% of the total clones).

The matched-pair test of symmetry found 11 pairwise comparisons out of 629 pairs between cloned sequences and their reference mtDNA counterparts to be statistically significant at α-level below 0.05, suggesting that

Table 2
Sequence characteristics of clones. Presence of stop codons was determined by translating the nucleotide into amino acid sequences. Presence of insertions and deletions was determined by comparing each clone with its corresponding mdDNA.

| Species | No of total clones | No of unique clones | No of clones with stop codon(s) | No of clones with insertion(s) | No of clones with deletion(s) |
|---|---|---|---|---|---|
| *S. albolineata* | 35 | 27 | 12 | 4 | 10 |
| *S. shoshone* | 49 | 12 | 2 | 2 | 2 |
| *S. ceratiola* | 48 | 36 | 9 | 4 | 7 |
| *S. alutacea* | 51 | 33 | 6 | 3 | 6 |
| *S. rubiginosa* | 22 | 20 | 9 | 6 | 6 |
| *S. obscura* | 40 | 16 | 2 | 0 | 2 |
| *S. lineata* | 41 | 20 | 4 | 0 | 4 |
| *S. impleta* | 50 | 40 | 7 | 2 | 6 |
| *S. damnifica* | 43 | 17 | 2 | 1 | 3 |
| *S. flavofasciata* | 51 | 23 | 6 | 4 | 7 |
| *S. quisqueya* | 54 | 19 | 1 | 1 | 2 |
| *S. vaga* | 52 | 31 | 17 | 12 | 14 |
| *S. americana* | 79 | 66 | 36 | 5 | 32 |
| *S. piceifrons* | 46 | 34 | 18 | 4 | 21 |
| *S. socorro* | 52 | 20 | 3 | 1 | 2 |
| *S. pallens* | 46 | 38 | 8 | 0 | 9 |
| *S. cancellata* | 36 | 35 | 13 | 5 | 18 |
| *S. gregaria* | 43 | 43 | 18 | 6 | 21 |
| *S. literosa* | 57 | 27 | 9 | 5 | 7 |
| *S. melanocera* | 50 | 22 | 5 | 3 | 7 |
| *H. robustus* | 56 | 50 | 27 | 2 | 20 |
| Total | 1001 | 629 | 214 | 70 | 206 |

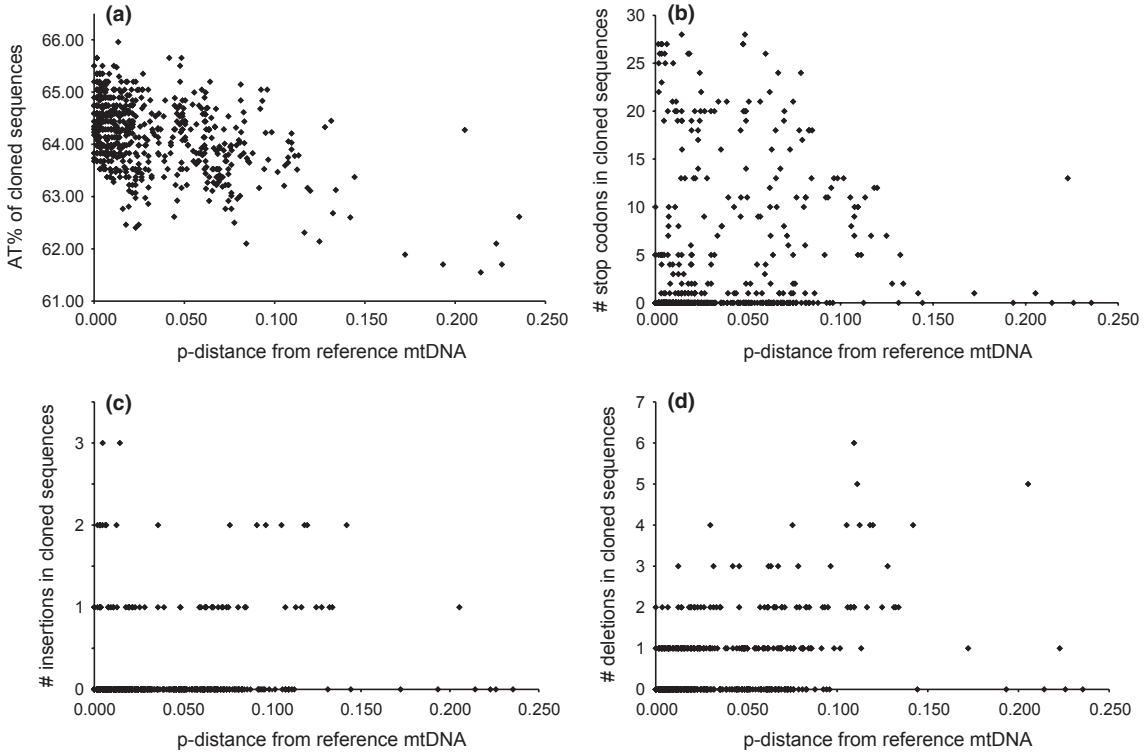


Fig. 2. Sequence characteristics of cloned sequences represented as scatterplots. (a) AT% against the uncorrected *p*-distance from the reference mtDNA; (b) number of stop codons against the uncorrected *p*-distance from the reference mtDNA; (c) number of insertions against the uncorrected *p*-distance from the reference mtDNA; (d) number of deletions against the uncorrected *p*-distance from the reference mtDNA.

the overall data were consistent with evolution under globally stationary, reversible, and homogeneous conditions (Table 3). A similar pattern was observed when AT% of all clones was plotted against uncorrected *p*-distances at the nucleotide level (Fig 2a). A majority of points were clumped at similar AT% with low sequence divergence, suggesting that most numts had the similar base compositions to the reference mtDNA.

We also found that a large number of clones did not accumulate either indels or stop codons (375 clones or 59.6% of the total clones). Among these clones, the sequence divergence relative to the reference mtDNA varied between 0% and 23.5%. Those clones with considerably divergent nucleotide sequences or very different base compositions from the reference mtDNA were clearly identified as numts. However, it was difficult to determine the numt status for those clones that were similar to the reference mtDNA without the characteristic mutations. The numt status of these remaining clones was determined by the subsequent phylogenetic analyses.

### *Phylogenetic analyses*

The parsimony analysis based on all unique clones and the reference mtDNA sequences (655-terminal dataset) found 292 equally parsimonious trees (length = 5075, CI = 0.27, RI = 0.79). The strict consensus collapsed 410 nodes and the relationships among different species were largely unresolved (Fig. 3). Nevertheless, the analysis recovered many monophyletic clades consisting of a reference mtDNA of a particular species and a majority of cloned sequences generated from the same species. There were numerous cases in which cloned sequences from different species formed monophyletic groups. Many clones that did not accumulate either indels or stop codons formed polytomy with the corresponding reference mtDNA sequences.

The parsimony analysis of the 26-terminal dataset based on the reference mtDNA sequences resulted in two equally parsimonious trees (length = 417, CI = 0.54, RI = 0.65). The four-gene dataset resulted in a single most parsimonious tree (length = 741, CI = 0.61, RI = 0.66). Both datasets produced nearly identical topologies (Fig. 4). The Bayesian analyses of these two datasets also resulted in similar topologies with high posterior probabilities (data not shown). The ingroups (*Schistocerca* and *Halmenus*) formed a strong monophyletic group, with *S. gregaria* as the basal lineage. The Galapagos endemic brachypterous *Halmenus* was recovered as sister to two Galapagos *Schistocerca* species (*S. literosa* and *S. melanocera*), and this clade represented the earliest diverging branch of the New World *Schistocerca*. The internal relation-

Table 3
Sequence characteristics of clones and the results from the matched-pairs test of symmetry. AT% was calculated to represent a base composition. Uncorrected *p*-distance was calculated in comparison with the reference mtDNA. The matched-pairs test of symmetry tests against the null hypothesis that a pair of sequences has evolved under the same stationary, reversible and homogeneous conditions. For our dataset, we performed this test between the reference mtDNA and unique cloned sequences for each species. If the test finds a pairwise comparison with a significant *P*-value, it would indicate that a cloned sequence in that comparison has evolved under a heterogeneous condition. We also show the smallest *P*-value obtained amongst the pairwise comparisons for each species for comparative purposes. Overall, 11 of 629 tests produced *P*-values below 0.05, implying that the data are consistent with evolution under globally stationary, reversible, and homogeneous conditions.

| Species | AT% mtDNA | AT% among clones | *p*-distance among clones | No of pairwise comparisons | No of pairs with *P*-value < 0.05 | smallest *P*-value amongst comparisons |
|---|---|---|---|---|---|---|
| *S. albolineata* | 65.35 | 62.60–65.96 | 0.002–0.142 | 27 | 2 | 0.0310 |
| *S. shoshone* | 65.20 | 61.70–65.50 | 0.000–0.226 | 12 | 0 | 0.0860 |
| *S. ceratiola* | 64.29 | 63.53–65.35 | 0.006–0.083 | 36 | 1 | 0.0302 |
| *S. alutacea* | 64.59 | 63.77–65.65 | 0.000–0.131 | 33 | 0 | 0.1573 |
| *S. rubiginosa* | 64.89 | 63.22–65.05 | 0.003–0.096 | 20 | 0 | 0.0611 |
| *S. obscura* | 64.29 | 63.98–64.59 | 0.002–0.079 | 16 | 0 | 0.1102 |
| *S. lineata* | 65.50 | 64.38–65.65 | 0.000–0.067 | 20 | 0 | 0.1116 |
| *S. impleta* | 64.89 | 61.55–65.35 | 0.000–0.214 | 40 | 1 | 0.0137 |
| *S. damnifica* | 64.74 | 62.14–65.65 | 0.000–0.125 | 17 | 1 | 0.0294 |
| *S. flavofasciata* | 64.74 | 63.37–65.05 | 0.000–0.144 | 23 | 0 | 0.1573 |
| *S. quisqueya* | 63.83 | 63.37–64.13 | 0.000–0.021 | 19 | 0 | 0.1573 |
| *S. vaga* | 64.29 | 62.31–65.05 | 0.000–0.116 | 31 | 0 | 0.0986 |
| *S. americana* | 63.98 | 62.40–64.49 | 0.002–0.205 | 66 | 0 | 0.0817 |
| *S. piceifrons* | 64.13 | 61.89–64.84 | 0.007–0.172 | 34 | 0 | 0.1116 |
| *S. socorro* | 64.44 | 63.47–64.59 | 0.000–0.024 | 20 | 1 | 0.0455 |
| *S. pallens* | 63.53 | 62.61–65.20 | 0.044–0.235 | 38 | 2 | 0.0361 |
| *S. cancellata* | 64.29 | 62.46–65.50 | 0.016–0.132 | 35 | 3 | 0.0025 |
| *S. gregaria* | 64.74 | 62.10–65.05 | 0.055–0.134 | 43 | 0 | 0.1131 |
| *S. literosa* | 63.83 | 62.96–65.05 | 0.000–0.120 | 27 | 0 | 0.0611 |
| *S. melanocera* | 64.29 | 63.21–64.44 | 0.000–0.105 | 22 | 0 | 0.0833 |
| *H. robustus* | 64.69 | 61.70–65.20 | 0.012–0.193 | 50 | 0 | 0.0863 |


Fig. 3. A strict consensus of 292 most parsimonious trees based on 655-terminal dataset including all unique clones and the reference mtDNA. Each branch is colour coded to indicate its species identity; terminals with yellow circles represent the reference mtDNA sequences. The presence of insertions (blue), deletions (green), and stop codons (orange) is mapped onto the terminals that have these mutations. While many cloned sequences have indels and stop codons, there are numerous clones without such mutations. Terminal labels are intentionally removed to reduce the complexity of the figure. Several clades consisting of the same-coloured branches (e.g. red, green, yellow, purple), one of which includes the reference mtDNA, are the clade consisting of the numts and the mtDNA of a single species. Those clades consisting of the differently coloured branches are the clades that are made up of synaponumts of different species. See Fig. 8 for more detailed example.

ships were largely congruent with the current classification, but supported with low nodal supports.

The simultaneous analyses of the cloned sequences of a given species and the reference mtDNA sequences resulted in patterns congruent with the proposed conceptual framework of numt accumulation. The resulting tree statistics for these analyses are listed in Table 4. In all 21 species, the phylogenetic placement of the cloned sequences helped determine their status as either autaponumts or synaponumts. As expected, the autaponumts formed a monophyletic clade with the reference mtDNA. While some autaponumts had the characteristic mutations (indels and stop codons), many autaponumts did not have them. Some synaponumts formed clades with the reference mtDNA of other species in some cases (e.g. *S. albolineata* and *S. ceratiola* in Fig. 5), while other synaponumts were simply placed outside the monophyletic clades formed by the reference mtDNA and the autaponumts, but did not relate to other reference mtDNA sequences (e.g. *S. socorro* and *S. quisqueya* in Fig. 6). The phylogenetic distribution of numts also reflected the phylogenetic placement of a given species. For example, the numt sequences of three Galapagos endemic species (*S. literosa*, *S. melanocera*, and *H. robustus* in Fig. 7) were recovered only at the base of the phylogeny, suggesting that these lineages might have branched off early during the diversification of *Schistocerca*. In most cases, the phylogenetic distribution of numts corresponded well with the phylogenetic placement of the particular species. One exception was the desert locust. Both mtDNA-only and four-gene trees strongly suggested that *S. gregaria* was the earliest diverging species within the genus (Fig. 4), which would indicate that its numts should not group with the mtDNA sequences of the species that diverged later. However, the numts of *S. gregaria* were found to be closely related to the mtDNA sequences of many species that supposedly diverged later in the phylogeny of *Schistocerca* (Fig. 7).

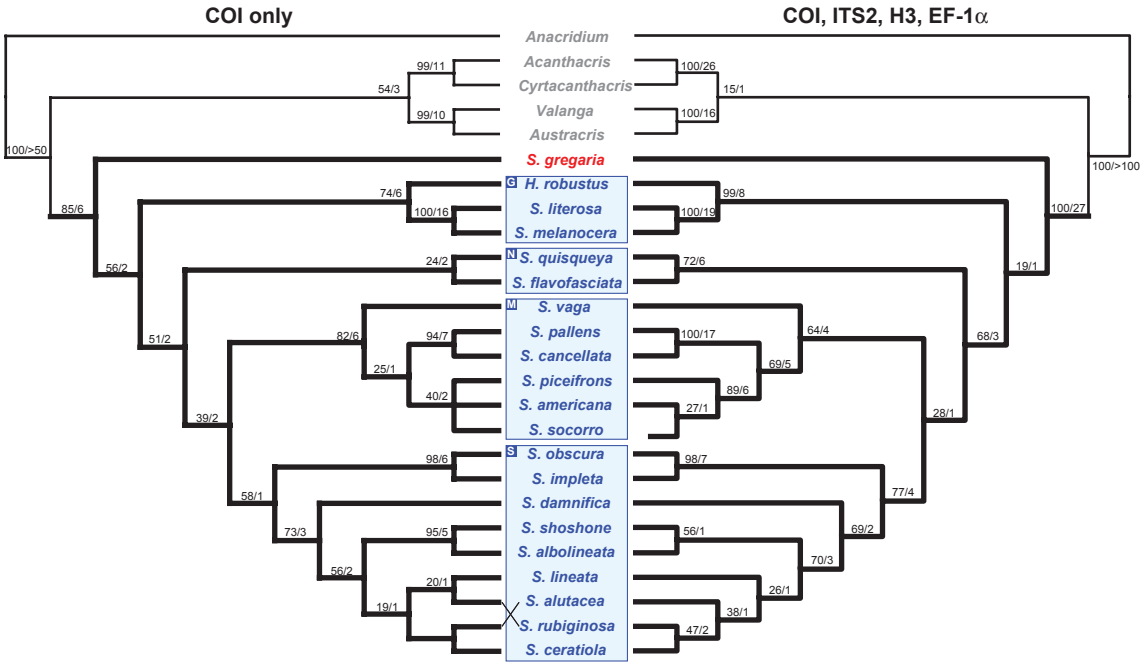


Fig. 4. A strict consensus of two most parsimonious trees based on COI gene (left) and a single most parsimonious tree based on COI, ITS, H3, and EF-1α gene (right). In both analyses, *Schistocerca* (+ *Halmenus*) are recovered as a strong monophyletic group, with *S. gregaria* at the base of the phylogeny. The Old World species is coloured red and the New World species are coloured blue. Among the New World species, four clades are identified as: G, Galapagos species; N, two early branching species in the continental America; M, the mobile clade *sensu* Song (2004a); S, the sedentary clade *sensu* Song (2004a). The nodal support values (bootstrap/Bremer) are shown above the nodes.

Table 4
Tree statistics from the simultaneous analyses of numts of a given species and the reference mtDNA of all species across *Schistocerca*. MPT refers to the most parsimonious tree. CI and RI are the ensemble consistency index and the ensemble retention index, respectively.

| Species | No of MPT | Length | CI | RI |
|---|---|---|---|---|
| *S. albolineata* | 40 | 636 | 0.52 | 0.69 |
| *S. shoshone* | 4 | 564 | 0.56 | 0.72 |
| *S. ceratiola* | 1301 | 692 | 0.49 | 0.69 |
| *S. alutacea* | 94 | 591 | 0.51 | 0.67 |
| *S. rubiginosa* | 502 | 569 | 0.52 | 0.67 |
| *S. obscura* | 23 | 509 | 0.52 | 0.71 |
| *S. lineata* | 96 | 508 | 0.52 | 0.67 |
| *S. impleta* | 208 | 755 | 0.52 | 0.66 |
| *S. damnifica* | 54 | 600 | 0.52 | 0.67 |
| *S. flavofasciata* | 1 | 733 | 0.54 | 0.70 |
| *S. quisqueya* | 58 | 466 | 0.54 | 0.74 |
| *S. vaga* | 14 | 689 | 0.52 | 0.71 |
| *S. americana* | 265 | 1164 | 0.50 | 0.67 |
| *S. piceifrons* | 444 | 862 | 0.51 | 0.63 |
| *S. socorro* | 1 | 482 | 0.52 | 0.72 |
| *S. pallens* | 617 | 764 | 0.51 | 0.71 |
| *S. cancellata* | 48 | 839 | 0.47 | 0.72 |
| *S. gregaria* | 189 | 1021 | 0.46 | 0.67 |
| *S. literosa* | 76 | 717 | 0.52 | 0.73 |
| *S. melanocera* | 12 | 619 | 0.53 | 0.72 |
| *H. robustus* | 1526 | 786 | 0.52 | 0.74 |

## Discussion

### *Numts are extremely prevalent within Schistocerca*

One of the major findings in our study is that numts are extremely prevalent in every single species of *Schistocerca* we examined (Table 2). Grasshoppers belonging to Acrididae are known to have a large number of numts (Bensasson et al., 2000), possibly due to very large nuclear genome sizes (Bensasson et al., 2001b; Hazkani-Covo, 2009) and the fact that numts tend to accumulate frequently in non-coding regions of the genome (Bensasson et al., 2000). Previous studies have shown that the prevalence of numts is lineage-specific when different subfamilies and families are compared (Song et al., 2008; Moulton et al., 2010), but the present study strongly demonstrates that closely related species in lower-level taxonomic groups (within a species complex or genus) could have similarly high accumulation of numts. In other words, the prevalence of numts in *Schistocerca* is a general feature for the genus, rather than an anomaly.

Although the mutation rate in numts is known to be much slower than that in the mitochondrial genes (Arctander, 1995; Lopez et al., 1997), it has been

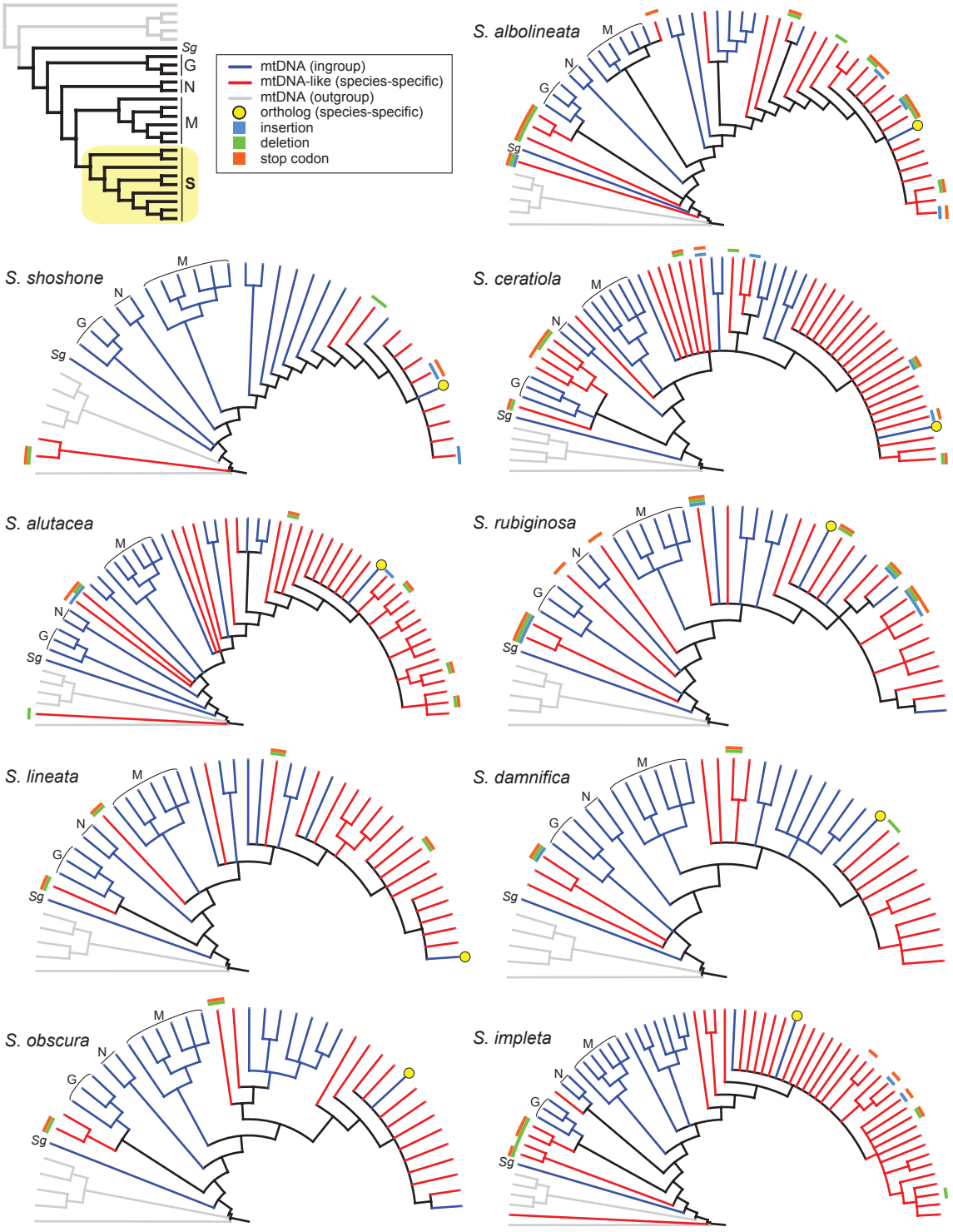


Fig. 5. Phylogenetic distribution of numts in the species in the sedentary clade based on simultaneous analyses of numts of a given species and the reference mtDNA of all species included in the study. For each species, either the strict consensus or the single most parsimonious tree is shown. Red branches indicate the numt sequences generated from a given species; blue branches indicate the reference mtDNA of the ingroup species. The mtDNA of the given species is indicated with a yellow circle. Grey branches indicate outgroup species. Among the ingroups, three New World clades identified in Fig. 4 (excluding the sedentary clade) are indicated on each tree with the letters G, N, and M. *Sg* indicates *S. gregaria*. The numts with characteristic mutations are shown with coloured boxes placed next to the terminals: insertion (light blue), deletion (green), and stop codon (orange). The phylogeny of *Schistocerca* based on four genes is shown at top left as a reference.

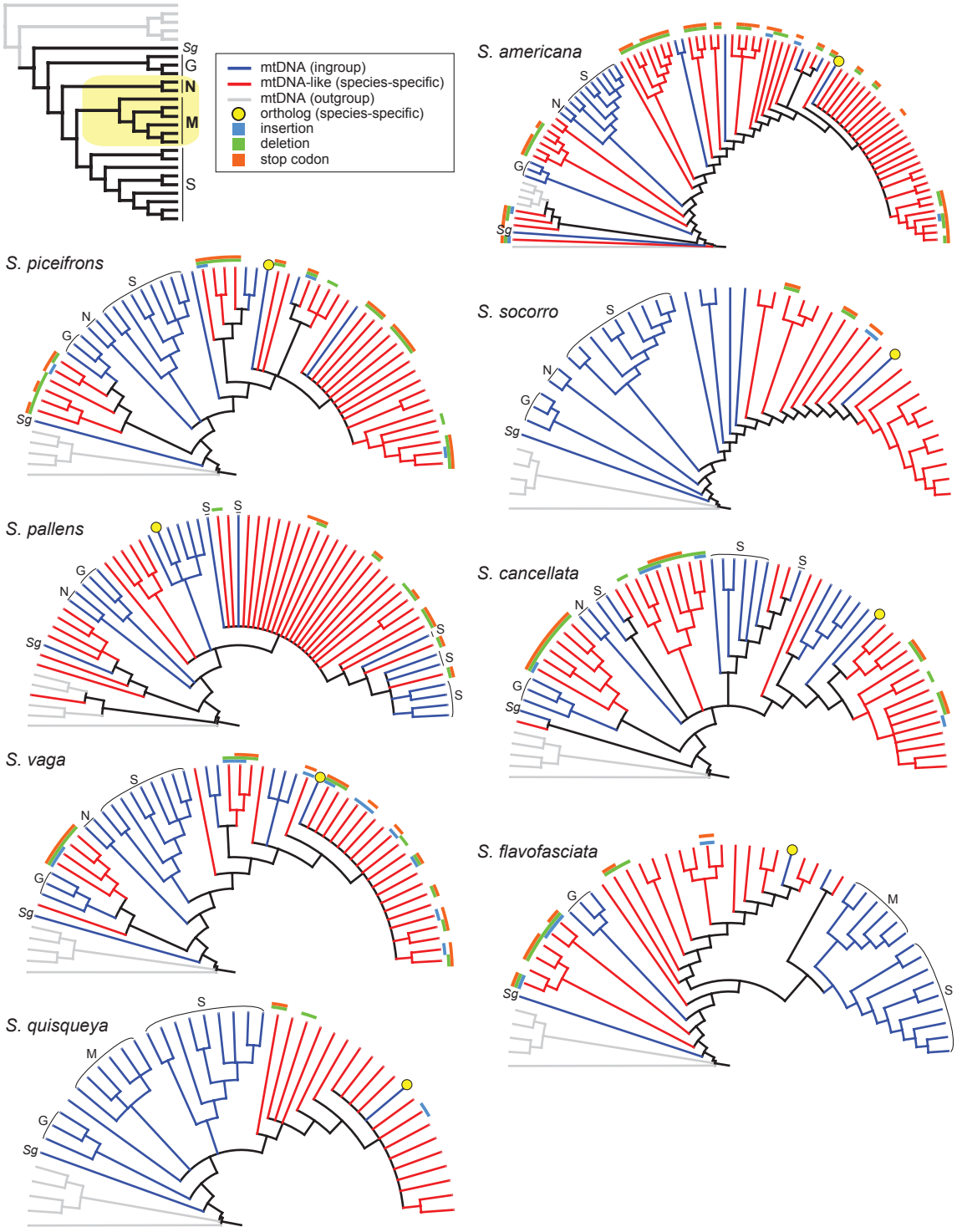


Fig. 6. Phylogenetic distribution of numts in the species in the mobile clade and the first continental clade based on simultaneous analyses of numts of a given species and the reference mtDNA of all species included in the study. For each species, either the strict consensus or the single most parsimonious tree is shown. Red branches indicate the numt sequences generated from a given species; blue branches indicate the reference mtDNA of the ingroup species. The mtDNA of the given species is indicated with a yellow circle. Grey branches indicate outgroup species. Among the ingroups, three New World clades identified in Fig. 4 are indicated on each tree with the letters G, N, M, and S. *Sg* indicates *S. gregaria*. The numts with characteristic mutations are shown with coloured boxes placed next to the terminals: insertion (light blue), deletion (green), and stop codon (orange). The phylogeny of *Schistocerca* based on four genes is shown at top left as a reference.

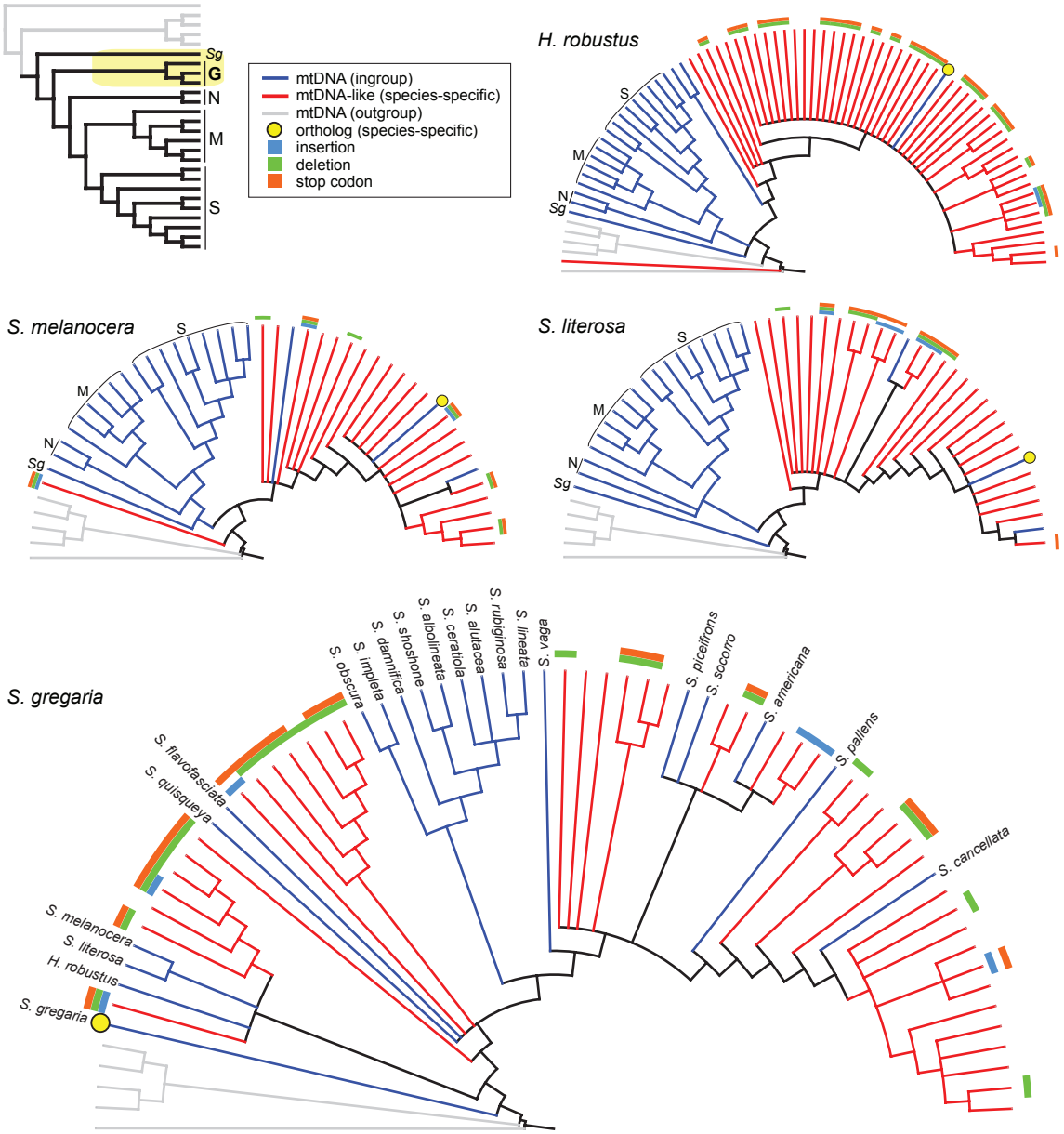


Fig. 7. Phylogenetic distribution of numts in the species in the Galapagos clade and *S. gregaria* based on simultaneous analyses of numts of a given species and the reference mtDNA of all species included in the study. For each species, either the strict consensus or the single most parsimonious tree is shown. Red branches indicate the numt sequences generated from a given species; blue branches indicate the reference mtDNA of the ingroup species. The mtDNA of the given species is indicated with a yellow circle. Grey branches indicate outgroup species. Among the ingroups, three New World clades identified in Fig. 4 are indicated on each tree with the letters N, M, and S. *Sg* indicates *S. gregaria*. The numts with characteristic mutations are shown with coloured boxes placed next to the terminals: insertion (light blue), deletion (green), and stop codon (orange). The phylogeny of *Schistocerca* based on four genes is shown at top left as a reference.

reported that numts do accumulate random mutations once integrated into nuclear genome (Bensasson et al., 2001a). As such, the signature characteristic of numts of protein-coding genes is the presence of indels and stop codons when compared with the extant mtDNA. This is a frequently observed pattern in our dataset, and about 40% of the cloned sequences can be identified easily as numts on this basis. However, we also find that nearly 60% of the clones do not possess either indels or stop codons. Instead, they simply have numerous point mutations that have been accumulated randomly and ostensibly appear to be functional. A more in-depth sequence characterization is necessary to determine their numt status.

Because the nuclear genome evolves under a different condition from the mitochondrial genome, numts that have been integrated into the nuclear genome for a long period may have distinct characteristics that can be

identified on careful inspection. First, whereas the base composition of the mitochondrial genome is typically biased toward A and T (Wolstenhome, 1992), it is less biased toward A and T in the nuclear genome, thus numts are expected to have base compositions distinctly different from the reference mtDNA (Bensasson et al., 2001a). Second, because random mutations may accumulate continuously, there would be considerable sequence divergence at the level of nucleotide between the numts and the reference mtDNA. However, our analysis based on the matched-pairs test of symmetry reveals that only a small proportion of the numts (11 out of 629, 1.7% of total clones) is determined to have statistically different base compositions from the reference mtDNA. On the other hand, the sequence divergence between the numts and the reference mtDNA appears to meet the expectation better. While nearly one-third of the clones have less than 1% sequence divergence from the mtDNA, many clones are highly divergent. The most extreme case is that of one cloned sequence from *S. pallens*, which is 23.5% divergent from the reference mtDNA without accumulating either indels or stop codons. The observed pattern appears to suggest that numts tend to accumulate random point mutations to result in considerable sequence divergence from the mtDNA, but it might take a very long time to accumulate enough mutations to have different base compositions. We suspect that our methodology might have influenced the observed pattern. Our data generation method relied on the idea that conserved primers would anneal to numts because the priming sites would still be similar enough for efficient PCR amplification. By doing so, we have effectively limited our numt exploration ability to finding only those numts that can be coamplified with the conserved primers. In other words, we have biased our ability to sampling only those numts that are similar enough to the mtDNA, and thus the coamplified numts are necessarily a subset of the diversity of numts in the nuclear genome. Theoretically, it makes sense for numts to have different base composition from the mtDNA and these numts would have accumulated numerous mutations in all sites including the priming sites. Our PCR-based methodology precluded the generation of those highly divergent numts, which might be the reason why only a very small proportion of the clones have statistically different base compositions from the mtDNA.

Interestingly, 89 clones, ca. 14% of the total, are difficult to characterize clearly as numts. All these have accumulated point mutations without indels and stop codons, and all the point mutations are silent substitutions, mostly in the third codon positions and occasionally in the first codon positions in the case of leucine. These clones may represent recent numts because it is possible that mtDNA would continue to evolve by changes at silent sites while numts would retain ancestral sequence of mtDNA (Bensasson et al., 2001a). It is also possible to consider that these silent substitutions are random point mutations that simply happen to accumulate at the third codon positions. Some of these clones differ from the reference mtDNA by only a few base pairs. It is possible that these clones may represent errors associated with PCR and cloning, but the chance of them being errors may be minimal at best because of the high-fidelity polymerase we used in data generation. Bertheau et al. (2011) reported numts of COI gene from the European bark beetle that were one to three base pairs different from the mtDNA sequence, and called them cryptic numts. Similarly, Berthier et al. (2011) found COI-like sequences that were only slightly different from the mtDNA in two locust species, and hypothesized that they would represent microheteroplasmic copies. In reality, the distinction between cryptic numts and microheteroplasmies is difficult to make, but the presence of mtDNA-like sequences that are only slightly different from the mtDNA is certainly widespread among different *Schistocerca* species and possibly in other taxa as well. However, a small proportion of these clones (20 clones, 3% of total) have too high of sequence divergence at the nucleotide level (1.1–6.3%) to be considered recent numts, cryptic numts, or microheteroplasmies. There may be two biologically sensible explanations for these clones, neither of which is easy to test. The first possibility is that these clones represent nuclear copies of mtDNA of other species that the extant species hybridized with in the past. Baldo et al. (2011) showed this pattern in several divergent populations of jumping bristletails, and it would be feasible for different species of *Schistocerca* to hybridize in nature (Harvey, 1979; Jago et al., 1979), especially because the genus appears to be of a recent origin (Song, 2004a). The second possibility is that these clones are highly divergent heteroplasmic copies. Heteroplasmy can be caused by somatic mutations within an individual, sometimes associated with slipped-strand mispairing during replication (Densmore et al., 1985; Lunt et al., 1998), aging (Lin et al., 2002), random genetic drift through the mitochondrial genetic bottleneck during the formation of the female germ line (Chinnery et al., 2000), or biparental inheritance of mitochondria (Hoeh et al., 1991) or paternal leakage (Kondo et al., 1990; Gyllensten et al., 1991; Kvist et al., 2003). Some heteroplasmic copies can be highly dissimilar from the prevalent copy (Kvist et al., 2003; White et al., 2008). Although the presence of heteroplasmy was not confirmed in *S. gregaria* by Bensasson et al. (2000), the possibility of this phenomenon occurring in *Schistocerca* should not be dismissed. Regardless of whether these clones are numts, divergent heteroplasmies, or numts of divergent heteroplasmies, our findings clearly add to the myriad of confounding factors in the use of mtDNA in systematics (Ballard and Whitlock, 2004; White et al., 2008; Triant and Hayes, 2011).

*Empirical support for the conceptual framework of numt accumulation*

Our study aims at testing the utility of the conceptual framework of numt accumulation (Fig. 1) and the possibility of using numts as phylogenetically informative markers systematically. By correctly characterizing each cloned sequence against the reference mtDNA and by phylogenetically analysing these sequences, we are able to find a pattern that is largely congruent with our prediction.

A large proportion of clones can be characterized as the autaponumts because they form a monophyletic group with the mtDNA, and this finding confirms that the nuclear integration of mtDNA is not an isolated one-time event in a species' evolutionary history, but an ongoing process that continues after species divergence (Fig. 8a). However, when these monophyletic clades containing both the autaponumts and the mtDNA are examined in detail, we find that the mtDNA is often placed as one of the apical tips nested within the clade, and rarely near the base of the clade. This pattern suggests that the extant mtDNA has continuously and rapidly mutated since the time of species divergence, and that we are able to capture a series of snapshots of ancestral mtDNA in the form of the autaponumts even though these numts must have also gone through random mutations. The relatively slower rate of mutation in numts is what makes them useful as molecular fossils. Some of these numts form clades among themselves, suggesting a possible duplication event or multiple nuclear integration events at the same time.

We also find several monophyletic clades consisting of synaponumts from different species (Fig. 8b,c), suggesting that the nuclear genomes of these species harbour the ancestral forms of mtDNA that were integrated prior to each species divergence, which have remained relatively intact despite randomly accumulating mutations that must have occurred continuously over time (Perna and Kocher, 1996; Hazkani-Covo, 2009). The presence of recognizable synaponumts is thus indicative of the shared ancestry at some point during the diversification of the lineage. The phylogenetic analysis of the 655-terminal dataset finds two types of synaponumts, differing in the time of divergence. The first type results from the nuclear integration in the most recent common ancestor. For example, the species belonging to the Americana Complex are closely related (Harvey, 1981), and the numts from closely related species such as *S. americana* and *S. piceifrons* often form clades indicating that the numts of the most recent common ancestor can be readily coamplified from the immediate descendant species (Fig. 8b). The second type results from the nuclear integration in the ancient common ancestors at various times. These clades of synaponumts from different species are often placed at the base of the phylogeny (Fig. 8c). We also find that the synaponumts from two or more species that do not share a recent common ancestor often form clades. One extreme example is a clade consisting of synaponumts from *Halmenus*, *S. shoshone*, *S. impleta*, *S. piceifrons*, and *S. pallens*, which do not share a recent common ancestor, and their most recent common ancestor would be the ancestral *Schistocerca* that gave rise to the current diversity.

It is possible to argue that characterizing numts into two types (autaponumts and synaponumts) resulting only from cladogenetic events might be overly simplistic. Although it is difficult to test directly, there are other evolutionary processes that can result in similar types of numts discovered in our study. For example, if two closely related species have frequently hybridized in the past, that has resulted in mitochondrial introgression, and if the fragments of introgressed mtDNA were integrated into the nuclear genome, it would be possible to coamplify numts from one species that would resemble mtDNA of another species. Two closely related species that have not completed lineage sorting at the level of mtDNA can harbour numts from other species in their nuclear genomes.

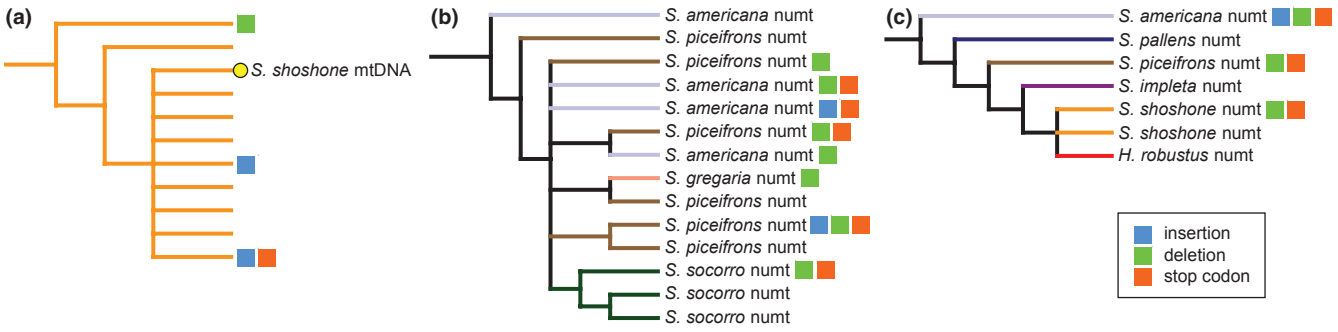


Fig. 8. Examples of (a) autaponumts, (b) synaponumts of closely related species, and (c) synaponuts of divergent species forming clades, taken from the 655-terminal tree. Each terminal is colour-coded to indicate species identity. The presence of insertions (light blue), deletions (green), and stop codons (orange) is mapped onto the terminal that has these characteristics.

Furthermore, these numts would likely form monophyletic clades with numts or mtDNA of other species when sequences from both species are analysed together (Baldo et al., 2011). Based on the conceptual framework presented in this study, however, these numts would be categorized as synaponumts, when in fact the evolutionary processes giving rise to these is not cladogenetic. Unfortunately, there is not a good way of distinguishing these two possibilities other than sampling many individuals from each species. However, our analysis found a number of clades consisting of numts from divergent species, not necessarily of numts from two closely related species. Regardless of whether these numts resulted from direct nuclear integration in the common ancestors or by past hybridization events, the distribution of numts opens up fascinating and complex evolutionary histories of these grasshoppers.

Incidentally, our conceptual framework developed for numt accumulation also has a potential implication for the study of horizontal gene transfer (HGT). HGT is a well known mechanism for the diversification and speciation in bacteria because genes from one species can be readily incorporated into the genome of another species via transformation, transduction, and conjugation (Ochman et al., 2000). Increasingly, there have been several reports of HGT of plant mitochondrial genes between parasitic plants and their hosts (Bergthorsson et al., 2004; Bock, 2010). Hao et al. (2010) found that in the plant genus *Ternstroemia* there were several species with mitochondrial genes that arose by conversion between native mitochondrial copies and foreign homologs from other species introduced by HGT. Numts have been reported from plants, sometimes with high frequencies in species such as *Arabidopsis thaliana*, *Oryza sativa*, and *Sorghum bicolor* (Hazkani-Covo et al., 2010). These findings collectively raise an intriguing possibility that some of the numts in plant genomes might carry historical signature of past HGT events, which could hint at ancient interspecific ecological interactions.

### *Phylogenetic distribution of numts sheds new light on the biogeography of Schistocerca*

One of the core ideas behind the conceptual framework of numt accumulation is that a given species would harbour both autaponumts and synaponumts in its nuclear genome that have been accumulated throughout all its cladogenetic events. Because the synaponumts in particular represent molecular fossils, or relatively "frozen" snapshots of the ancient mtDNA throughout the species' divergence, it is possible that the synaponumts of a given species may group with the extant mtDNA of another species if two species shared a common ancestor at any time throughout their evolution. Therefore a simultaneous phylogenetic analysis of the numts from one species and the mtDNA of closely related species can reveal intricate details about their evolution, not readily visible from a typical phylogenetic study.

*Schistocerca* is one of two New World genera of a predominantly Old World grasshopper subfamily Cyrtacanthacridinae (Song and Wenzel, 2008). The other New World genus is *Halmenus*, a small brachypterous genus endemic to the Galapagos Islands. The close relationship between *Schistocerca* and *Halmenus* has been supported consistently by morphological characters (Dirsh, 1974; Amédégnato, 1993; Song, 2004a; Song and Wenzel, 2008), but Lovejoy et al. (2006) found that *Halmenus* was closely related to two Galapagos endemic *Schistocerca* species (*S. literosa* and *S. melanocera*), suggesting a possibility that *Halmenus* might be a divergent and brachypterous member of *Schistocerca* that had been incorrectly classified as a different genus. Our analysis based on four genes strongly corroborates Lovejoy et al.'s (2006) finding, in that these three Galapagos species are recovered as a monophyletic group (Clade G in Fig. 4), placed at the base of the New World *Schistocerca* (Fig. 4). This means that the Galapagos Islands were colonized by the common ancestor of these three species and that *Halmenus* is in fact *Schistocerca* that lost its wings possibly due to insular brachypterism (Amédégnato, 1993). The numt-based analyses add an interesting temporal aspect to the divergence of these three species (Fig. 7). The synaponumts of *S. melanocera* form clades with the mtDNA of *S. literosa* and *Halmenus*. Likewise, the synaponumts of *S. literosa* form clades with the mtDNA of *S. melanocera* and *Halmenus*. However, the synaponumts of *Halmenus* do not form clades with either *S. literosa* or *S. melanocera*. This pattern suggests that *Halmenus* must have diverged much earlier than the speciation events that led to *S. literosa* and *S. melanocera* and therefore the nuclear genome of *Halmenus* never had an opportunity to accumulate the synaponumts with the other Galapagos species, according to the conceptual framework of numt accumulation. In a similar vein, the synaponumts of these three Galapagos species do not form clades with the mtDNA of other New World *Schistocerca* species, suggesting that the Galapagos clade is the earliest diverging lineage within the New World *Schistocerca*. Considering when the Galapagos Islands became available for terrestrial colonization (3–4 Ma, Peck, 2001), *Schistocerca* as a whole appears to be a very young genus.

Among the remaining New World species, a clade consisting of *S. flavofasciata* and *S. quisqueya* branches off first (Clade N in Fig. 4). Their synaponumts are distributed near their corresponding mtDNA or at the base of the phylogeny and do not

form clades with other continental species (Fig. 6), suggesting that these two species must have diverged prior to the major radiation of *Schistocerca* in North and South America. The remaining *Schistocerca* is divided into two major clades, which we refer to here as the mobile clade and the sedentary clade, following Song (2004a). The mobile clade includes *S. vaga*, *S. pallens*, *S. cancellata*, *S. americana*, *S. piceifrons*, and *S. socorro* (clade M in Fig. 4), which roughly corresponds to the Americana Complex *sensu* Harvey (1981). The Americana Complex includes a group of morphologically similar species that have strong flight capacity, and Harvey (1981) considered all swarming locust species (*S. gregaria*, *S. piceifrons*, and *S. cancellata*) as a part of this complex. However, our study finds *S. gregaria* placed at the base of the phylogeny (Fig. 4), which suggests that the morphological similarities among the species in the complex must be due to convergent evolution. The sedentary clade comprises the North American species including *S. obscura*, *S. impleta*, *S. damnifica*, *S. shoshone*, *S. albolineata*, *S. lineata*, *S. alutacea*, *S. rubiginosa*, and *S. ceratiola* (clade S in Fig. 4). These species, although also strong fliers, are ecologically sedentary and often display host plant and habitat preference associated with host plants (Sword and Chapman, 1994; Sword and Dopman, 1999). Song (2004b) recovered six species (*S. alutacea*, *S. rubiginosa*, *S. lineata*, *S. shoshone*, *S. albolineata*, and *S. obscura*) as a monophyletic group based on morphology and called them the Alutacea Group, but our four-gene analysis recovers the Alutacea Group as paraphyletic (Fig. 4). The phylogenetic distribution of synaponumts of the species belonging to the mobile and sedentary clades reveal interesting evolutionary dynamics not evident from the species phylogeny alone. The synaponumts of the several species in the mobile clade are distributed near the mtDNA of the mobile clade or at the base of the phylogeny, but do not group with the mtDNA of the sedentary clade (Fig. 6). Similarly, the synaponumts from the most species in the sedentary clade do not group with the mtDNA of the mobile clade (Fig. 5). This pattern suggests that the major cladogenetic event splitting the mobile and sedentary clades must have occurred early in the process of diversification of *Schistocerca*, and the two clades remained separate since the initial divergence. However, the phylogenetic distribution of synaponumts of *S. cancellata* and *S. pallens*, which belong to the mobile clade, does not follow this pattern in that some of these synaponumts group with the mtDNA of the sedentary clade (Fig. 6). A close inspection of these numts reveals that they are highly divergent from their corresponding mtDNA and possibly represent numts of divergent heteroplasmies, which may have yielded the unexpected placement of these numts. The fact that *S. cancellata* and *S. pallens* are sister species seems to suggest that perhaps both species shared similar divergent heteroplasmies.

Regarding the origin of the desert locust *S. gregaria*, our four-gene analysis clearly points to the Old World origin because of its basal placement in the phylogeny of *Schistocerca* (Fig. 4). This finding corroborates with Lovejoy et al.'s (2006) study, but is in conflict with the morphological phylogeny by Song (2004a). In light of the new data, it appears that many of the morphological characters that Song (2004a) used to group *S. gregaria* with the species in the Americana Complex must have been homoplasious at best. According to the conceptual framework of numt accumulation, the numts of an early diverging lineage, such as *S. gregaria*, should be recovered only at the base of the phylogeny in a simultaneous analysis of numts and mtDNA (Fig. 7). However, we find that the numts of *S. gregaria* group with the mtDNA of the Galapagos species, *S. flavofasciata* and *S. quisqueya*, as well as the members of the mobile clade, but not with the sedentary clade. The phylogenetic distribution of *S. gregaria* numts (Fig. 7) is eerily similar to that of other species in the mobile clade (Fig. 6), which suggests that the correct phylogenetic placement of *S. gregaria* may be actually in the mobile clade. In other words, the numts of *S. gregaria* carry historical signature of the New World origin, which is in direct conflict with the pattern deduced from the four-gene analysis. How can this discrepancy be explained?

In explaining the origin of *S. gregaria* there have been three alternative hypotheses, two of which are the New World origin hypothesis and the Old World origin hypothesis (Song, 2004a). The third hypothesis is a variant of the Old World origin hypothesis known as the Multiple Crossings hypothesis, which states that the New World *Schistocerca* species could have been a result of multiple colonization events by the ancestral *gregaria*-like species from the Old World. Kevan (1989) initially suggested this hypothesis soon after a spectacular incident in which a large swarm of desert locust successfully crossed the Atlantic Ocean from the West Africa to reach the West Indies in October 1988 (Kevan, 1989; Ritchie and Pedgley, 1989). This hypothesis was later expanded by Amédégnato (1993), but did not gain a wide acceptance and was later dismissed by both Song (2004a) and Lovejoy et al. (2006) because it would require an explanation for multiple extinction events of the several founding species that originated from Africa. If *S. gregaria* were the source species that gave rise to the New World *Schistocerca* via multiple westward transatlantic colonization events at different times, its nuclear genome should contain numerous numts of ancestral mtDNA representing different ancient colonizers. These numts would then be

similar to the mtDNA of those species that descended from the particular colonizer, which is exactly what we observe. Interestingly, the numts of *S. gregaria* do not group with the mtDNA of the sedentary clade (Fig. 7), which suggests that the divergence of this clade must have been an actual cladogenetic event confined in the New World that did not involve another colonization event from Africa. Jago et al. (1979) hybridized three New World *Schistocerca* species (now known as *S. americana*, *S. cancellata*, and *S. pallens*) with the African *S. gregaria*. They found that in most cases the hybridization experiments produced over 90% non-viable eggs with no developing embryos. However, the cross between *S. pallens* females and *S. gregaria* males produced about 14% viable eggs, while the cross between *S. americana* females and *S. gregaria* males produced about 52% viable eggs. If the New World *Schistocerca* species were the result of single colonization by the ancestral *S. gregaria*, we would expect that hybridization between *S. gregaria* and the New World species would be impossible simply because the amount of morphological divergence that must have accumulated throughout the divergence of the genus. However, Jago et al.'s (1979) findings provide evidence that perhaps the divergence between *S. gregaria* and the New World species is not as great, consistent with the idea that there might have been a more recent colonization event from Africa. Therefore our study based on numts raises a distinct possibility that the New World *Schistocerca* may be a result of multiple colonization events from the African *S. gregaria*, which is otherwise a very challenging hypothesis to test based on traditional characters.

### *Concluding remarks*

Numts clearly present a challenge for molecular systematics using mtDNA (Zhang and Hewitt, 1996; Sorenson and Quinn, 1998; Bensasson et al., 2001a; Birky, 2001; Ballard and Whitlock, 2004; Sword et al., 2007; Song et al., 2008; White et al., 2008). However, given that we can reliably identify numts, they can be a valuable source of information because they represent the molecular fossils of the mtDNA. Analogous to the limitations of paleontological data that are affected by physical changes on Earth, numts are limited to the extent that they are affected by random mutations in the nuclear genome. Nevertheless, when viewed as molecular fossils and ancestral mtDNA, numts provide interesting and powerful phylogenetic signal, much more than what extant mtDNA or nuclear gene sequences might be able to provide. We acknowledge that the pattern found in our study may not be directly applicable to other organisms because a system like ours requires a recently diverged lineage with numerous numt accumulations throughout its diversification. However, as numts are reported from more and more organisms (Hazkani-Covo et al., 2010), there will be more opportunities to discover systems similar to that presented in this study.

## Acknowledgements

D. Otte, L. Barrientos-Lozano, S.B. Peck, S.J. Simpson, B. Valentine, D. Perez-Gelabert, late R.F. Chapman, the members of the Whiting Lab kindly contributed specimens used in this analysis. We thank K.A. Crandall and J.W. Sites for valuable discussion on how to analyse numts, N.C. Sheffield for writing a script to characterize numts, and S.L. Cameron and G.A. Sword for providing critical comments on the earlier version of the manuscript. J. Sullivan, L. Jermiin, D. Morrison, D. Bensasson, R. DeBry, and three anonymous reviewers provided insightful criticism on an earlier manuscript. This work was supported by the National Science Foundation (DEB-1064082 to H.S. and EF-0531665 to M.F.W.).

## References

Ababneh, F., Jermiin, L.S., Ma, C., Robinson, J., 2006. Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. Bioinformatics 22, 1225–1231.

Amédégnato, C. 1993. African–American relationships in the Acridians (Insecta, Orthoptera). In: George, W., Lavocat, R. (Eds.), The Africa–South America Connection. Clarendon Press, Oxford, pp. 59–75.

Arctander, P., 1995. Comparison of a mitochondrial gene and a corresponding nuclear pseudogene. Proc. R. Soc. Lond. B Biol. Sci. 262, 13–19.

Baldo, L., de Queiroz, A., Hedin, M., Hayashi, C.Y., Gatesy, J., 2011. Nuclear-mitochondrial sequences as witnesses of past interbreeding and population diversity in the jumping bristletail *Mesomachilis*. Mol. Biol. Evol. 28, 195–210.

Ballard, J.W.O., Whitlock, M.C., 2004. The incomplete natural history of mitochondria. Mol. Ecol. 13, 729–744.

Behura, S.K., 2007. Analysis of nuclear copies of mitochondrial sequences in honeybee (*Apis mellifera*) genome. Mol. Biol. Evol. 24, 1492–1505.

Benesh, D.P., Hasu, T., Suomalainen, L.R., Valtonen, E.T., Tiirola, M., 2006. Reliability of mitochondrial DNA in an acanthocephalan: the problem of pseudogenes. Int. J. Parasitol. 36, 247–254.

Bensasson, D., Zhang, D.-X., Hewitt, G.M., 2000. Frequent assimilation of mitochondrial DNA by grasshopper nuclear genomes. Mol. Biol. Evol. 17, 406–415.

Bensasson, D., Zhang, D.-X., Hartl, D.L., Hewitt, G.M., 2001a. Mitochondrial pseudogenes: evolution's misplaced witnesses. Trends Ecol. Evol. 16, 314–321.

Bensasson, D., Petrov, D.A., Zhang, D.-X., Hartl, D.L., Hewitt, G.M., 2001b. Genomic gigantism: DNA loss is slow in mountain grasshoppers. Mol. Biol. Evol. 18, 246–253.

Bensasson, D., Feldman, M.W., Petrov, D.A., 2003. Rates of DNA duplication and mitochondrial DNA insertion in the human genome. J. Mol. Evol. 57, 343–354.

Bergthorsson, U., Richardson, A.O., Young, G.J., Goertzen, L.R., Palmer, J.D., 2004. Massive horizontal transfer of mitochondrial

genes from diverse land plant donors to the basal angiosperm *Amborella*. Proc. Natl Acad. Sci. USA 101, 17747–17752.
Bertheau, C., Schuler, H., Krumbock, S., Arthofer, W., Stauffer, C., 2011. Hit or miss in phylogeographic analyses: the case of the cryptic NUMTs. Mol. Ecol. Resour. 11, 1056–1059.
Berthier, K., Chapuis, M.-P., Moosavi, S.M., Tohidi-Esfahani, D., Sword, G.A., 2011. Nuclear insertions and heteroplasmy of mitochondrial DNA as two sources of intra-individual genomic variation in grasshoppers. Syst. Entomol. 36, 285–299.
Birky, C.W., 2001. The inheritance of genes in mitochondria and chroloplasts: laws, mechanisms, and models. Annu. Rev. Genet. 35, 125–148.
Bock, R., 2010. The give-and-take of DNA: horizontal gene transfer in plants. Trends Plant Sci. 15, 11–22.
Brown, W.M., Prager, E.M., Wang, A., Wilson, A.C., 1982. Mitochondrial DNA sequences of primates: tempo and mode of evolution. J. Mol. Evol. 18, 225–239.
Buhay, J.E., 2009. "COI-like" sequences are becoming problematic in molecular systematic and DNA barcoding studies. J. Crustac. Biol. 29, 96–110.
Chinnery, P.F., Thorburn, D.R., Samuels, D.C., White, S.L., Dahl, H.M., Turnbull, D.M., Lightowlers, R.N., Howell, N., 2000. The inheritance of mitochondrial DNA heteroplasmy: random drift, selection or both? Trends Genet. 16, 500–505.
Collins, F.H., Paskewitz, S.M., 1996. A review of the use of ribosomal DNA (rDNA) to differentiate among cryptic *Anopheles* species. Insect Mol. Biol. 5, 1–9.
Collura, R.V., Stewart, C.B., 1995. Insertions and duplications of mtDNA in the nuclear genomes of Old World monkeys and hominoids. Nature 378, 485–489.
Densmore, L.D., Wright, J.W., Brown, W.M., 1985. Length variation and heteroplasmy are frequent in mitochondrial DNA from parthenogenetic and bisexual lizards (genus *Cnemidophorus*). Genetics 110, 689–707.
Dirsh, V.M. 1974. Genus *Schistocerca* (Acridomorpha, Insecta). Dr W. Junk, The Hague.
Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797.
Erler, S., Ferenz, H.-J., Moritz, R.F.A., Kaatz, H.-H., 2010. Analysis of the mitochondrial genome of *Schistocerca gregaria gregaria* (Orthoptera: Acrididae). Biol. J. Linn. Soc. 99, 296–305.
Folmer, O., Black, M., Hoeh, W., Lutz, R., Vrijenhoek, R., 1994. DNA primers for amplification of mitochondrial cytochrome *c* oxidase subunit I from diverse metazoan invertebrates. Mol. Mar. Biol. Biotechnol. 3, 294–299.
Fukuda, M., Wakasugi, S., Tsuzuki, T., Nomiyama, H., Shimada, K., Miyata, T., 1985. Mitochondrial DNA-like sequences in the human nuclear genome: characterization and implications in the evolution of mitochondrial DNA. J. Mol. Biol. 186, 257–266.
Goloboff, P.A., 1999. Analyzing large data sets in reasonable times: solutions for composite optima. Cladistics 15, 415–428.
Goloboff, P.A., Farris, J.S., Nixon, K.C. 2003. T.N.T.: Tree Analysis Using New Technology. Program and documentation. Available from the authors and at www.zmuc.dk/public/phylogeny.
Gyllensten, U., Wharton, D., Josefsson, A., Wilson, A.C., 1991. Parental inheritance of mitochondrial DNA in mice. Nature 352, 255–257.
Hao, W., Richardson, A.O., Zheng, Y., Palmer, J.D., 2010. Gorgeous mosaic of mitochondrial genes created by horizontal transfer and gene conversion. Proc. Natl. Acad. Sci. U. S. A. 107, 21576–21581.
Harvey, A.W., 1979. Hybridization studies in the *Schistocerca americana* complex I. The specific status of the Central American locust. Biol. J. Linn. Soc. 12, 349–355.
Harvey, A.W., 1981. A reclassification of the *Schistocerca americana* complex (Orthoptera: Acrididae). Acrida 10, 61–77.
Hay, J.M., Sarre, S.D., Daugherty, C.H., 2004. Nuclear mitochondrial pseudogenes as molecular outgroups for phylogenetically isolated taxa: a case study in *Sphenodon*. Heredity 93, 468–475.
Hazkani-Covo, E., 2009. Mitochondrial insertions into primate nuclear genomes suggest the use of numts as a tool for phylogeny. Mol. Biol. Evol. 26, 2175–2179.
Hazkani-Covo, E., Sorek, R., Graur, D., 2003. Evolutionary dynamics of large numts in the human genome: rarity of independent insertions and abundance of post-insertion duplications. J. Mol. Evol. 56, 169–174.
Hazkani-Covo, E., Zeller, R.M., Martin, W., 2010. Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes. PLoS Genet. 6, e1000834.
Hlaing, T., Tun-Lin, W., Somboon, P., Socheat, D., Setha, T., Min, S., Chang, M.S., Walton, C., 2009. Mitochondrial pseudogenes in the nuclear genome of *Aedes aegypti* mosquitoes: implications for past and future population genetic studies. BMC Genet. 10, 11.
Ho, J.W.K., Adams, C.E., Lew, J.B., Matthews, T.J., Ng, C.C., Shahabi-Sirjani, A., Tan, L.H., Zhao, Y., Easteal, S., Wilson, S.R., Jermiin, L.S., 2006. SeqVis: visualization of compositional heterogeneity in large alignments of nucleotides. Bioinformatics 22, 2162–2163.
Hoeh, W.R., Blakley, K.H., Brown, W.M., 1991. Heteroplasmy suggests limited biparental inheritance of *Mytilus* mitochondrial DNA. Science 251, 1488–1490.
Huson, D.H., Richter, D.C., Rausch, C., Dezulian, T., Franz, M., Rupp, R., 2007. Dendroscope – an interactive viewer for large phylogenetic trees. BMC Bioinformatics 8, 460.
Jago, N.D., Antonious, A., Scott, P., 1979. Laboratory evidence showing the separate species status of *Schistocerca gregaria*, *americana* and *cancellata* (Acrididae, Cyrtacanthacridinae). Syst. Entomol. 4, 133–142.
Jensen-Seaman, M.I., Wildschutte, J.H., Soto-Calderon, I.D., Anthony, N.M., 2009. A comparative approach shows differences in patterns of numt insertion during hominoid evolution. J. Mol. Evol. 68, 688–699.
Kevan, D.K.M., 1989. Transatlantic travelers. Antenna 13, 12–15.
Knowles, L.L., Carstens, B.C., 2007. Delimiting species without monophyletic gene trees. Syst. Biol. 56, 887–895.
Kondo, R., Satta, Y., Matsuura, E.T., Ishiwa, H., Takahata, N., Chigusa, S.I., 1990. Incomplete maternal transmission of mitocondrial DNA in *Drosophila*. Genetics 126, 657–663.
van der Kuyl, A.C., Kuiken, C.L., Dekker, J.T., Perizonius, W.R., Goudsmit, J., 1995. Nuclear counterparts of the cytoplasmic mitochondrial 12S rRNA gene: a problem of ancient DNA and molecular phylogenies. J. Mol. Evol. 40, 625–657.
Kvist, L., Martens, J., Nazarenko, A.A., Orell, M., 2003. Paternal leakage of mitochondrial DNA in the great tit (*Parus major*). Mol. Biol. Evol. 20, 243–247.
Leroux, C., Issel, C.J., Montelaro, R.C., 1997. Novel and dynamic evolution of equine infectious anemia virus genomic quasispecies associated with sequential disease cycles in an experimentally infected pony. J. Virol. 71, 9627–9639.
Letunic, I., Bork, P., 2007. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics 23, 127–128.
Letunic, I., Bork, P., 2011. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. Nucleic Acids Res. 39 (suppl 2), W475–W478.
Lin, M.T., Simon, D.K., Ahn, C.H., Kim, L.M., Beal, M.F., 2002. High aggregate burden of somatic mtDNA point mutations in aging and Alzheimer's disease brain. Hum. Mol. Genet. 11, 133–145.
Lopez, J.V., Yuhki, N., Masuda, R., Modi, W., O'Brien, S.J., 1994. Numt, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat. J. Mol. Evol. 39, 174–190.
Lopez, J.V., Culver, M., Stephens, J.C., Johnson, W.E., O'Brien, S.J., 1997. Rates of nuclear and cytoplasmic mitochondrial DNA sequence divergence in mammals. Mol. Biol. Evol. 14, 277–286.
Lovejoy, N.R., Mullen, S.P., Sword, G.A., Chapman, R.F., Harrison, R.G., 2006. Ancient trans-Atlantic flight explains

locust biogeography: molecular phylogenetics of *Schistocerca*. Proc. R. Soc. Lond. B 273, 767–774.
Lunt, D.H., Whipple, L.E., Hyman, B.C., 1998. Mitochondrial DNA variable number tandem repeats (VNTRs): utility and problems in molecular ecology. Mol. Ecol. 7, 1441–1455.
Maddison, W.P., Knowles, L.L., 2006. Inferring phylogeny despite incomplete lineage sorting. Syst. Biol. 55, 21–30.
Maddison, D.R., Maddison, W.P. 2005. MacClade 4. Sinauer Associates, Sunderland, MA.
Miraldo, A., Hewitt, G.M., Dear, P.H., Paulo, O.S., Emerson, B.C., 2012. Numts help to reconstruct the demographic history of the ocellated lizard (*Lacerta lepida*) in a secondary contact zone. Mol. Ecol. 21, 1005–1018.
Mirol, P.M., Mascheretti, S., Searle, J.B., 2000. Multiple nuclear pseudogenes of mitochondrial cytochrome *b* in *Ctenomys* (Caviomorpha, Rodentia) with either great similarity to or high divergence from the true mitochondrial sequence. Heredity 84, 538–547.
Moulton, M.J., Song, H., Whiting, M.F., 2010. Assessing the effects of primer specificity on eliminating numt coamplification in DNA barcoding: a case study from Orthoptera (Arthropoda: Insecta). Mol. Ecol. Resour. 10, 615–627.
Nixon, K.C., 1999. The parsimony ratchet, a new method for rapid parsimony analysis. Cladistics 15, 407–414.
Nylander, J.A.A. 2004. MrModeltest v.2. Program distributed by the author, Evolutionary Biology Centre, Uppsala University.
Ochman, H., Lawrence, J.G., Groisman, E.A., 2000. Lateral gene transfer and the nature of bacterial innovation. Nature 405, 299–304.
Peck, S.B. 2001. Smaller orders of insects of the Galápagos Islands, Ecuador: evolution, ecology, and diversity. NRC Research Press, Ottawa.
Perna, N.T., Kocher, T.D., 1996. Mitochondrial DNA: molecular fossils in the nucleus. Curr. Biol. 6, 128–129.
Podnar, M., Haring, E., Pinsker, W., Mayer, W., 2007. Unusual origin of a nuclear pseudogene in the Italian Wall Lizard: intergenomic and interspecific transfer of a large section of the mitochondrial genome in the genus *Podacris* (Lacertidae). J. Mol. Evol. 64, 308–320.
Richly, E., Leister, D., 2004. NUMTs in sequenced eukaryotic genomes. Mol. Biol. Evol. 21, 1081–1084.
Ritchie, M., Pedgley, D.E., 1989. Desert locusts cross the Atlantic. Antenna 13, 10–12.
Ronquist, F., Huelsenbeck, J.P., 2003. MRBAYES 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19, 1572–1574.
Ruokonen, M., Kvist, L., Lumme, J., 2000. Close relatedness between mitochondrial DNA from seven *Anser* goose species. J. Evol. Biol. 13, 532–540.
Schmitz, J., Piskurek, O., Zischler, H., 2005. Forty million years of independent evolution: a mitochondrial gene and its corresponding nuclear pseudogenes in primates. J. Mol. Evol. 61, 1–11.
Song, H., 2004a. On the origin of the desert locust *Schistocerca gregaria* (Forskål) (Orthoptera: Acrididae: Cyrtacanthacridinae). Proc. R. Soc. Lond. B Biol. Sci. 271, 1641–1648.
Song, H., 2004b. Revision of the alutacea group of genus *Schistocerca* (Orthoptera: Acrididae: Cyrtacanthacridinae). Ann. Entomol. Soc. Am. 97, 420–436.
Song, H., Wenzel, J.W., 2008. Phylogeny of bird–grasshopper subfamily Cyrtacanthacridinae (Orthoptera: Acrididae) and the evolution of locust phase polyphenism. Cladistics 24, 515–542.
Song, H., Buhay, J.E., Whiting, M.F., Crandall, K.A., 2008. Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. Proc. Natl Acad. Sci. USA 105, 13486–13491.
Sorenson, M.D., Quinn, T.W., 1998. Numts: a challenge for avian systematics and population biology. Auk 115, 214–221.
Sunnucks, P., Hales, D.F., 1996. Numerous transposed sequences of mitochondrial cytochrome oxidase I–II in aphids of the genus *Sitobion* (Hemiptera: Aphididae). Mol. Biol. Evol. 13, 510–524.
Svenson, G.J., Whiting, M.F., 2004. Phylogeny of Mantodea based on molecular data: evolution of a charismatic predator. Syst. Entomol. 29, 359–370.
Sword, G.A., Chapman, R.F., 1994. Monophagy in a polyphagous grasshopper, *Schistocerca shoshone*. Entomol. Exp. Appl. 73, 255–264.
Sword, G.A., Dopman, E.B., 1999. Developmental specialization and geographic structure of host plant use in a polyphagous grasshopper, *Schistocerca emarginata* (= *lineata*) (Orthoptera: Acrididae). Oecologia 120, 437–445.
Sword, G.A., Senior, L.B., Gaskin, J.F., Joern, A., 2007. Double trouble for grasshopper molecular systematics: intra-individual heterogeneity of both mitochondrial 12S-valine-16S and nuclear internal transcribed spacer ribosomal DNA sequences in *Hesperotettix viridis* (Orthoptera: Acrididae). Syst. Entomol. 32, 420–428.
Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., Kumar, S., 2011. MEGA5: Molecular Evolutionary Genetics Analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. 28, 2731–2739.
Thalmann, O., Hebler, J., Poinar, H.N., Pääbo, S., Vigilant, L., 2004. Unreliable mtDNA data due to nuclear insertions: a cautionary tale from analysis of humans and other great apes. Mol. Ecol. 13, 321–335.
Thalmann, O., Serre, D., Hofreiter, M., Lukas, D., Eriksson, J., Vigilant, L., 2005. Nuclear insertions help and hinder inference of the evolutionary history of gorilla mtDNA. Mol. Ecol. 14, 179–188.
Triant, D.A., DeWoody, J.A., 2007. Extensive mitochondrial DNA transfer in a rapidly evolving rodent has been mediated by independent insertion events and by duplications. Gene 401, 61–70.
Triant, D.A., Hayes, L.D., 2011. Molecular approaches in behavioural research: a cautionary note regarding mitochondrial transfers to the nucleus (numts). Anim. Behav. 82, 601–606.
White, D.J., Wolff, J.N., Pierson, M., Gemmell, N.J., 2008. Revealing the hidden complexities of mtDNA inheritance. Mol. Ecol. 17, 4925–4942.
Williams, S.T., Knowlton, N., 2001. Mitochondrial pseudogenes are pervasive and often insidious in the snapping shrimp genus *Alpheus*. Mol. Biol. Evol. 18, 1484–1493.
Wolstenhome, D.R., 1992. Animal mitochondrial DNA: structure and evolution. Int. Rev. Cytol. 141, 173–216.
Zhang, D.-X., Hewitt, G.M., 1996. Nuclear integrations: challenge for mitochondrial DNA markers. Trends Ecol. Evol. 11, 247–251.
Zischler, H., 2000. Nuclear integrations of mitochondrial DNA in primates: inference of associated mutational events. Electrophoresis 21, 531–536.
Zischler, H., Geisert, H., von Haeseler, A., Pääbo, S., 1995. A nuclear "fossil" of the mitochondrial D-loop and the origin of modern humans. Nature 378, 489–492.